

Assessor burden, inter-rater agreement and user experience of the RoB-SPEO tool for assessing risk of bias in studies estimating prevalence of exposure to occupational risk factors: An analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury

Natalie C. Momen, Kai N. Streicher, Denise T. C. da Silva, Alexis Descatha, Monique H. W. Frings-Dresen, Diana Gagliardi, Lode Godderis, Tom Loney, Daniele Mandrioli, Alberto Modenese, et al.

▶ **To cite this version:**

Natalie C. Momen, Kai N. Streicher, Denise T. C. da Silva, Alexis Descatha, Monique H. W. Frings-Dresen, et al.. Assessor burden, inter-rater agreement and user experience of the RoB-SPEO tool for assessing risk of bias in studies estimating prevalence of exposure to occupational risk factors: An analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International, 2022, 158, 10.1016/j.envint.2021.107005 . hal-03511988

HAL Id: hal-03511988
https://hal.science/hal-03511988

Submitted on 5 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Environment International 158 (2022) 107005

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Assessor burden, inter-rater agreement and user experience of the RoB-SPEO tool for assessing risk of bias in studies estimating prevalence of exposure to occupational risk factors: An analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury

Natalie C. Momen [a], Kai N. Streicher [a], Denise T.C. da Silva [b], Alexis Descatha [c,d,e,f], Monique H. W. Frings-Dresen [g], Diana Gagliardi [h], Lode Godderis [i,j], Tom Loney [k], Daniele Mandrioli [l], Alberto Modenese [m], Rebecca L. Morgan [n], Daniela Pachito [o,p], Paul T.J. Scheepers [q], Daria Sgargi [l], Marília Silva Paulo [r], Vivi Schlünssen [s,t], Grace Sembajwe [u], Kathrine Sørensen [t], Liliane R. Teixeira [b], Thomas Tenkate [v], Frank Pega [a,*]

[a] Department of Environment, Climate Change and Health, World Health Organization, Geneva, Switzerland
[b] Workers' Health and Human Ecology Research Center, National School of Public Health Sergio Arouca, Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil
[c] UNIV Angers, CHU Angers, Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, Angers, France
[d] AP-HP (Paris Hospital), Occupational Health Unit, Poincaré University Hospital, Garches, France
[e] Versailles St-Quentin Univ-Paris Saclay Univ (UVSQ), UMS 011, UMR-S 1168, France
[f] Inserm, U1168 UMS 011, Villejuif, France
[g] Amsterdam UMC, University of Amsterdam, Department Public and Occupational Health/Coronel Institute of Occupational Health, Amsterdam Research Institute, Amsterdam, the Netherlands
[h] Inail, Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, Rome, Italy
[i] Centre for Environment and Health, KU Leuven, Leuven, Belgium
[j] KIR Department (Knowledge, Information & Research), IDEWE, External Service for Prevention and Protection at Work, Leuven, Belgium
[k] College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates
[l] Cesare Maltoni Cancer Research Center, Ramazzini Institute, Bologna, Italy
[m] Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy
[n] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[o] Núcleo de Avaliação de Tecnologias em Saúde, Hospital Sírio-Libanês, Bela Vista, São Paulo, SP, Brazil
[p] Fundação Getúlio Vargas, Bela Vista, São Paulo, SP, Brazil
[q] Radboud Institute for Health Sciences, Radboudumc, Nijmegen, the Netherlands
[r] Institute of Public Health, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates
[s] Aarhus University, Aarhus, Denmark
[t] National Research Center for the Working Environment, Copenhagen, Denmark
[u] Department of Environmental, Occupational, and Geospatial Health Sciences, CUNY Graduate School of Public Health and Health Policy, CUNY Institute for Implementation Science in Population Health, New York, NY, United States
[v] School of Occupational and Public Health, Ryerson University, Toronto, ON, Canada

ARTICLE INFO

Handling editor: Dr Paul Whaley

Keywords:
Bias
Systematic review methods

ABSTRACT

Background: As part of the development of the World Health Organization (WHO)/International Labour Organization (ILO) Joint Estimates of the Work-related Burden of Disease and Injury, WHO and ILO carried out several systematic reviews to determine the prevalence of exposure to selected occupational risk factors. Risk of bias assessment for individual studies is a critical step of a systematic review. No tool existed for assessing the risk of bias in prevalence studies of exposure to occupational risk factors, so WHO and ILO developed and pilot tested

* Corresponding author at: Department of Environment, Climate Change and Health, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland.
E-mail addresses: momenn@who.int (N.C. Momen), streicherk@who.int (K.N. Streicher), alexis.descatha@inserm.fr (A. Descatha), m.frings@amsterdamumc.nl (M.H.W. Frings-Dresen), d.gagliardi@inail.it (D. Gagliardi), lode.godderis@med.kuleuven.be (L. Godderis), tom.loney@mbru.ac.ae (T. Loney), mandriolid@ramazzini.it (D. Mandrioli), alberto.modenese@unimore.it (A. Modenese), morganrl@mcmaster.ca (R.L. Morgan), pachito@uol.com.br (D. Pachito), Paul.Scheepers@radboudumc.nl (P.T.J. Scheepers), mariliap@uaeu.ac.ae (M.S. Paulo), vs@ph.au.dk (V. Schlünssen), Grace.Sembajwe@sph.cuny.edu (G. Sembajwe), ksn@nfa.dk (K. Sørensen), lilianeteixeira@ensp.fiocruz.br (L.R. Teixeira), thomas.tenkate@ryerson.ca (T. Tenkate), pegaf@who.int (F. Pega).

https://doi.org/10.1016/j.envint.2021.107005
Received 7 April 2021; Received in revised form 19 November 2021; Accepted 23 November 2021
Available online 30 November 2021
0160-4120/© 2021 World Health Organization. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND IGO license (http://creativecommons.org/licenses/by-nc-nd/3.0/igo/).

Prevalence
Occupational exposure
Occupational epidemiology

the RoB-SPEO tool for this purpose. Here, we investigate the assessor burden, inter-rater agreement, and user experience of this new instrument, based on the abovementioned WHO/ILO systematic reviews.

Methods: Twenty-seven individual experts applied RoB-SPEO to assess risk of bias. Four systematic reviews provided a total of 283 individual assessments, carried out for 137 studies. For each study, two or more assessors independently assessed risk of bias across the eight RoB-SPEO domains selecting one of RoB-SPEO's six ratings (i.e., "low", "probably low", "probably high", "high", "unclear" or "cannot be determined"). Assessors were asked to report time taken (i.e. indicator of assessor burden) to complete each assessment and describe their user experience. To gauge assessor burden, we calculated the median and inter-quartile range of times taken per individual risk of bias assessment. To assess inter-rater reliability, we calculated a raw measure of inter-rater agreement (P_i) for each RoB-SPEO domain, between $P_i = 0.00$, indicating no agreement and $P_i = 1.00$, indicating perfect agreement. As subgroup analyses, P_i was also disaggregated by systematic review, assessor experience with RoB-SPEO ($\leq$10 assessments versus $>$ 10 assessments), and assessment time (tertiles: $\leq$25 min versus 26–66 min versus $\geq$ 67 min). To describe user experience, we synthesised the assessors' comments and recommendations.

Results: Assessors reported a median of 40 min to complete one assessment (interquartile range 21–120 min). For all domains, raw inter-rater agreement ranged from 0.54 to 0.82. Agreement varied by systematic review and assessor experience with RoB-SPEO between domains, and increased with increasing assessment time. A small number of users recommended further development of instructions for selected RoB-SPEO domains, especially bias in selection of participants into the study (domain 1) and bias due to differences in numerator and denominator (domain 7).

Discussion: Overall, our results indicated good agreement across the eight domains of the RoB-SPEO tool. The median assessment time was comparable to that of other risk of bias tools, indicating comparable assessor burden. However, there was considerable variation in time taken to complete assessments. Additional time spent on assessments may improve inter-rater agreement. Further development of the RoB-SPEO tool could focus on refining instructions for selected RoB-SPEO domains and additional testing to assess agreement for different topic areas and with a wider range of assessors from different research backgrounds.

1. Background

The World Health Organization (WHO) and the International Labour Organization (ILO) have produced the first WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury (WHO/ILO Joint Estimates) (Pega et al., 2021a,b, World Health Organization and International Labour Organization, 2021a,b). For this, WHO and ILO, along with a large number of individual experts, have conducted a series of systematic reviews providing the evidence base for these global health estimates (Li et al., 2018; Descatha et al., 2018; Godderis et al., 2018; Bonafede et al., 2021; Descatha et al., 2020; Hulshof et al., 2019; Li et al., 2020; Mandrioli et al., 2018; Pachito et al., 2021; Paulo et al., 2019; Pega et al., 2020a; Teixeira et al., 2019, 2021b; Tenkate et al., 2019; Hulshof et al., 2021a; Hulshof et al., 2021b; Rugulies et al., 2019; Teixeira et al., 2021a; Schlünssen et al., in preparation). An overview of the entire series is provided elsewhere (Pega et al., 2021c). Of these, five systematic reviews synthesised evidence from studies estimating the *prevalence* (or, in short, prevalence studies) of exposure to selected occupational risk factors. Prevalence of exposure is defined as the presence (and sometimes also level or dose) of a risk factor to human health (or a health outcome) among individuals within the study populations or a representative sample at one point in time (Porta, 2014). The systematic reviews aimed to determine the prevalence of exposure to five diverse occupational risk factors: ergonomic risk factors, dusts and/or fibres, solar ultraviolet radiation, noise, and long working hours. Such prevalence studies differ in several ways from studies that estimate incidence or prognosis, or those that estimate the effect of an exposure to an occupational risk factor on a health outcome (Pega et al., 2020b). There are differences in the type of data used (e.g., studying incidence requires longitudinal data, whereas studying prevalence utilises either longitudinal or cross-sectional data). Exposure prevalence studies focus solely on exposures, not health outcomes or estimation of effects. Furthermore, prevalence studies on occupational exposures only use data from studies of human subjects, whereas evidence on effects of exposures can also come from other evidence streams, namely mechanistic or animal data.

1.1. Risk of bias assessment of prevalence studies

A critical part of a systematic review is the assessment of the risk of bias (RoB) at the level of each included study. Risk of bias is the risk of "a systematic error, or deviation from the truth, in results" (Porta, 2014). When the development of the WHO/ILO Joint Estimates started, due to the differences between exposure prevalence studies and other study types that were highlighted above, no tool existed for assessing the RoB in prevalence studies of exposure to occupational risk factors (Krauth, Woodruff and Bero, 2013; Mandrioli and Silbergeld, 2016; Vandenberg et al., 2016; Whaley et al., 2016; NHMRC, 2019). None of the existing methods or instruments (Rooney et al., 2016; Morgan et al., 2019; Morgan et al., 2018b) were considered applicable for assessing prevalence studies of exposure to occupational risk factors (Pega et al., 2020b).

Two checklists exist for assessing RoB in individual prevalence studies (Hoy et al., 2012; Munn et al., 2014; Munn et al., 2015; The Joanna Briggs Institute, 2017), however, these checklists do not allow for recording of transparent rationales for ratings. Provision of reasons for assigned ratings is important as RoB assessment is based on judgment. Additionally, they also produce a summary score for RoB, which is controversial (Boutron et al., 2020).

Unlike the types of studies that existing methods focus on, occupational exposure prevalence studies do not consider health outcomes or effects. Additionally, as detailed above, they make use of different types of data. Therefore, WHO and ILO, supported by a large number of individual experts, developed the Risk of Bias in Studies estimating Prevalence of Exposure to Occupational risk factors (RoB-SPEO; (Pega et al., 2020b) tool, as a product of the WHO/ILO Joint Estimates. This new tool fills this gap and can be used in systematic reviews of occupational exposure prevalence studies. While drawing from existing tools, it is tailored to focus solely on domains relevant to assessing risk of bias in prevalence studies of occupational exposures.

1.2. The RoB-SPEO tool

The RoB-SPEO tool is described in detail in Pega et al., 2020b. Briefly, no prior existing tool for RoB assessment could be applied in its entirety to assess RoB in occupational exposure prevalence studies.

While components of prior existing tools may be applicable, they generally require revisions to make use of them in RoB assessments for studies of prevalence of occupational exposures because of key differences in focus and data used (as described above) between prevalence studies of exposure versus studies related to health outcomes and the effects of occupational exposures. Prevalence studies also have some unique biases (e.g. bias due to differences in numerator and denominator (Williams, Najman and Clavarino, 2006)), which are not comprehensively covered in methods in the prior existing tools for assessing RoB in other types of studies.

RoB-SPEO comprises eight domains that address different types of bias relevant to prevalence studies of exposure, as detailed in Table 1 (Pega et al., 2020b). Each domain consists of five components: i) guiding question to prompt the assessor; ii) description of bias and/or definitions of key terms and concepts; iii) considerations, which highlight potentially relevant domain-specific issues to assessors when conducting their assessment; iv) ratings and rating criteria and; v) a table for recording the assessment. For each study and each domain, RoB is judged using one of six standard ratings: i) low, ii) probably low, iii) probably high, iv) high, v) unclear and vi) cannot be determined. The tool was developed over multiple versions. Version 4.0 was pilot tested (Pega et al., 2020b), after which improvements were made (Appendix 1). Version 6.0 of RoB-SPEO was used to assess RoB in studies included in the five WHO/ILO Joint Estimates systematic reviews of prevalence.

Table 1
Domains of risk of bias in RoB-SPEO.

	Domain	Description of bias
1	Bias in selection of participants into the study	Bias in selection of participants into the study (commonly called selection bias) is the bias due to systematic differences between the characteristics of the study sample (defined as the sample of individuals participating in the study) and those of the target population (defined as the population for which the authors of the study sought to assess exposure) (Porta, 2014).
2	Bias due to lack of blinding of study personnel	Bias due to a lack of blinding of study personnel (commonly called performance bias) is the bias that arises when there is a lack of blinding of exposure assessors and other study personnel to relevant participant characteristics (e.g. disease status) that leads to exposure assessment that differs depending on participant characteristics.
3	Bias due to exposure misclassification	Bias due to exposure misclassification is "erroneous [and systematic] classification of an individual, a value, or an attribute into a [exposure] category other than that to which it should be assigned", leading to under- or over-estimation of prevalence of exposure status (or level) (Porta, 2014).
4	Bias due to incomplete exposure data	Bias due to incomplete exposure data is the bias that arise from exposure data missing in a way that the exposure assessment is differential by exposure status (or level) in the target population (i.e., not random).
5	Bias due to selective reporting of exposures	Bias due to selective exposure reporting is the systematic difference arising from selective reporting (under- or over-reporting) of exposures or exposure categories.
6	Bias due to conflict of interest	Bias due to conflicts of interest is the bias introduced if financial and other interests influence the design, conduct, data collection, analysis and/or reporting of a study (Woodruff and Sutton, 2014).
7	Bias due to differences in numerator and denominator	Bias due to differences in numerator and denominator is the bias that arises when there is a mismatch of definition and/or counting of persons contributing to the numerator and the denominator in the ratio used to estimate prevalence (Williams, Najman and Clavarino, 2006).
8	Other bias	Other bias is any other bias specific to a particular study rather than applicable to all studies.

Source: Pega et al., 2020b.

1.3. Assessor burden and inter-rater reliability of risk of bias tools

Ideally, a RoB tool should be efficient, placing the least possible burden on the assessor, while at the same time ensuring that an assessment is comprehensive and transparent. Therefore, studies testing the performance of RoB tools often assess the burden of the tool on assessors. A common indicator to that end is assessor-time per individual RoB assessment, generally measured in minutes (e.g., Jeyaraman et al., 2020a).

For a RoB tool to be fit for purpose, it must be a valid and reliable measure of the systematic differences from the truth. The tool's validity has been assessed extensively during its development and piloting (Pega et al., 2020b), but assessors' experience with the tool can provide further insights and are reported below. One measure of reliability is inter-rater reliability: the extent of agreement among raters (McHugh, 2012). Multiple assessors should apply the tool to assess a study, after which they should discuss conflicting ratings to reach consensus on final ratings. Individual judgement is an important part of RoB assessment; this makes perfect agreement for all studies an infeasible goal. In particular, it might be harder to reach agreement for studies which do not comprehensively report methods, data or results. However, ideally, variability in ratings between users should be low; detailed instructions for assessors, guiding them in how to reach a rating decision can help achieve this.

Several studies have assessed inter-rater reliability of domain-based RoB tools. In their protocol, Jeyaraman et al. (2020a) described that they plan to assess inter-rater reliability of their ROBINS-E instrument for non-randomized studies of the effect of environmental exposures on health outcomes using the AC1 statistic developed by Gwet (Gwet, 2001; Gwet, 2008). Assessments of other tools have calculated correlation (kappa) scores; this included assessments of the Evidence Project RoB tool (Kennedy et al., 2019), Cochrane Risk of Bias tool (Armijo-Olivo et al., 2012; Hartling et al., 2013), OHAT, IRIS, TSCA (Eick et al., 2020), Cochrane ROB 2.0 (Minozzi et al., 2020) and ROBINS-I tools (Couto et al., 2015). Similarly, interrater-reliability of the RoB-SPEO Version 4.0 was assessed using raw measures of agreement in the pilot testing (Pega et al., 2020b). While the improvements made on RoB-SPEO Version 4.0 were made with the main aim of aiding assessors to use the tool, it was hoped that while making it easier to use, the changes would also improve inter-rater agreement. We are not aware of any other inter-rater reliability assessments for RoB tools for environmental and occupational health, such as the RoB tools of the Navigation Guide (Woodruff and Sutton, 2014) and the US NIEHS Cancer Reports (NTP (National Toxicology Program), 2016), or the Risk of Bias Instrument for Non-randomized Studies of Exposures (Morgan et al., 2019).

1.4. Assessor burden and inter-rater agreement of RoB-SPEO in the pilot testing

During the pilot-testing phase of Version 4.0 of the RoB-SPEO tool, measurement of assessor burden was not reported, so this information has been unavailable for the tool so far. However, a raw measure of inter-rater agreement was calculated, as described previously (Pega et al., 2020b). Briefly, ratings were extracted by pilot tester, study domain, and record. The six standard RoB-SPEO ratings were coded into three analytical categories: i) low/probably low, ii) high/probably high, and iii) unclear/cannot be determined. During the piloting phase, the following levels of agreement were observed:

- Bias in selection of participants into the study: 0.33
- Bias due to a lack of blinding of study personnel: 0.65

- Bias due to exposure misclassification: 0.76
- Bias due to incomplete exposure data: 0.31
- Bias due to selective reporting of exposures: 0.80
- Bias due to conflict of interest: 0.51
- Other bias: 0.51 (Pega et al., 2020b)

Following this assessment during pilot testing, several changes were made to the RoB-SPEO tool, and it underwent an additional round of feedback and innovation, resulting in Version 6.0 of the RoB-SPEO (Appendix 1). Here, we describe the assessment of inter-rater agreement and assessor burden for this latest version of RoB-SPEO and present the results. While there are other methods and metrics for assessing inter-rater reliability, this opportunistic assessment uses data already generated as part of the series of systematic reviews for the WHO/ILO Joint Estimates. We believe that we present one of the first comprehensive assessments of the real-world performance of a tool for assessing RoB in environmental and occupational health studies, as applied in a global health policy setting.

2. Methods

Version 6.0 of the RoB-SPEO tool was applied by individual experts in exposure science and occupational and environmental health research participating in the WHO/ILO Joint Estimates systematic review series. They assessed study records for RoB across four out of five systematic reviews on prevalence of exposure in the series (Table 2). Two systematic reviews are completed (Hulshof et al., 2021a; 2021b; Teixeira et al., 2021a; 2021b), and three are ongoing (protocols: Descatha et al., 2018; Mandrioli et al., 2018; Paulo et al., 2019; Schlünssen et al., in preparation). One ongoing systematic review did not provide data for this assessment; it has not completed study selection at the time of writing, and therefore could not be included in this study. The complete data set of anonymised study records with their RoB-SPEO ratings can be accessed from Appendix 2.

2.1. Data

The data were collected as part of the four included systematic reviews conducted for the WHO/ILO Joint Estimates. A total of 342 risk of bias assessments made using RoB-SPEO version 6.0 and of studies included in the WHO/ILO systematic reviews were returned to WHO on request; 283 of the assessments made by 27 assessors for 137 studies were included in this examination of assessor burden and inter-rater reliability. Reasons for exclusion of studies and assessments are shown in Table 2.

2.1.1. Assessor burden

Assessors recorded the time (in minutes) spent completing each individual RoB assessment for each study. Overall, 23 assessors provided at least one time recording; time taken was stated for a total of 271 RoB assessments (Table 2). Time was missing for 12 RoB assessments made by nine assessors.

2.1.2. Inter-rater agreement

Each individual expert independently assessed each study record along each of the eight domains in the RoB-SPEO tool. For each domain, each assessor judged the RoB by assigning one of the six standard RoB-SPEO ratings: “low”, “probably low”, “probably high”, “high”, “unclear” or “cannot be determined”. Per study, each assessor recorded the selected RoB-SPEO rating for each RoB-SPEO domain in an individual assessment sheet. If an assessor provided no rating or a non-standard rating for a domain, these were treated as missing for the respective domain and excluded from the analysis for inter-rater agreement for the domain. Overall, 27 assessors provided a total of 2,249 RoB-SPEO ratings for a total of 283 assessments (Table 2). For domains 1, 5, 7 and 8, there were 1, 1, 6 and 7 ratings missing, respectively. Within domains, they all related to different studies resulting in the exclusion of the same number of studies from the analysis of each domain.

Table 2
Information about risk of bias assessments in this examination of assessor burden and inter-rater agreement.

Systematic review topic	Publications	No. of assessors who provided RoB assessments[a]	No. of studies included in each SR	No. of studies and assessments included and excluded
The prevalence of occupational exposure to ergonomic risk factors	Protocol and systematic review: Hulshof et al., 2019; Hulshof et al., 2021a	5	5	Included: 6 assessments for 3 studies Excluded: 2 studies • Only one RoB-SPEO assessment returned – 1 study[b] • No RoB-SPEO assessment returned – 1 study
The prevalence of occupational exposure to silica, asbestos and coal dust	Protocol and systematic review: Mandrioli et al., 2018; Schlünssen et al., in preparation	10	88	Included: 116 assessments for 54 studies Excluded: 34 studies • Assessments made at study record level – 28 assessments for 8 studies[c] • Only one RoB-SPEO assessment returned – 22 studies[b] • No RoB-SPEO assessment returned – 4 studies
The prevalence of occupational exposure to solar ultraviolet radiation	Protocol: Paulo et al., 2019	6	41	Included: 63 assessments for 31 studies Excluded: 10 studies • Only one RoB-SPEO assessment returned – 8 studies[b] • No RoB-SPEO assessment returned – 2 studies
The prevalence of occupational exposure to noise	Protocol and systematic review: Teixeira et al., 2019; Teixeira et al., 2021a	8	65	Included: 98 assessments for 49 studies Excluded: No RoB-SPEO assessment returned – 16 studies

[a] Two individuals were assessors for two systematic reviews; hence there were a total of 27 individual assessors across the four systematic reviews.
[b] Only receiving one RoB-SPEO assessment for a study meant that it was not possible to assess inter-rater agreement.
[c] Assessments were made at the study record level, not the study level.

2.1.3. User experience

As part of a survey of user experience, each assessor was asked to answer the following questions related to each domain:

- "What are the advantages of the new tool in the [relevant] domain?"
- "What are the disadvantages of the new tool in the [relevant] domain?"
- "How could the tool be further improved in the [relevant] domain?"

Additionally, they were asked about the performance of the entire tool:

- "What are the overall advantages of the new tool?"
- "What are the overall disadvantages of the new tool?"
- "How could the tool be further improved in general?"
- "Were any domains missing or could any domains be considered unnecessary?"

The responses were free text, providing qualitative data for analysis. In total, 13 assessors of studies included in this analysis responded to one or more of these questions.

2.2. Analysis

All analyses were conducted solely by authors of this current study (NCM, KNS, FP) who were not involved in the RoB assessments for the systematic reviews included in this study. The statistical analyses were conducted solely by authors who had not been involved in the development and pilot testing of RoB-SPEO (NCM, KNS).

2.2.1. Calculation of assessor burden

We calculated the median time taken (minutes) to complete one individual assessment and the interquartile range (due to the data being skewed; skewness values of −1.42 to −2.74 for each of the domains).

2.2.2. Calculation of interrater agreement

As was done for the pilot testing (Pega et al., 2020b), one author (NCM) coded the six RoB-SPEO ratings assigned by the assessors into the same three analytical categories: i) low/probably low, ii) high/probably high, and iii) unclear/cannot be determined. If an assessor provided no rating or a non-standard rating for a domain, these were treated as missing for the respective domain; the relevant study was then excluded from the analysis for inter-rater agreement for the domain. A second author (FP) independently checked the raw data against these codes. Random numbers were assigned to study records and individual assessors to ensure anonymity.

Separately for each of the eight RoB-SPEO domains, we calculated inter-rater agreement (P_i) across all assessors and all study records. The following formula (Morgan et al., 2018a; Armijo-Olivo et al., 2012; Bilandzic et al., 2016; Savovic et al., 2014; Losilla et al., 2018) was used to calculate the proportion of all ratings given by all pilot testers to the j-th analytical category (P_i):

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}\left(n_{ij}-1\right)$$

where $i = 1,\ldots d$ is the number of domains (here, $d = 8$); $j = 1,\ldots k$ is the number of possible analytical categories (here, $k = 3$); and n = number of assessors for the study record. Agreement could range from 0.00 (no two pilot testers chose the same rating) to 1.00 (pilot testers in all pairs chose the same rating).

In addition to examining overall agreement between assessors by domain, we carried out the following three subgroup analyses to assess the robustness of our results. First, we calculated inter-rater agreement by systematic review (anonymised and reported in random order). Second, we classified assessors by their experience with RoB-SPEO, classifying assessors who had carried out ≤ 10 included assessments using RoB as "less experienced" and those that assessed > 10 as "more experienced". We then calculated the inter-rater agreement by assessor experience, including in this subgroup analysis only those pairs of assessors whose experience with the tool was concordant, namely either both assessors carried out ≤ 10 studies or both carried our > 10 studies). Third, we divided time taken to complete the RoB assessment into tertiles (divided at the 33rd and 66th percentiles of reported times, i.e., short 0–25, medium 26–66, and long ≥ 67 min). If there was concordance between the times reported by the assessors on a study, the study was included in the calculation of inter-rater agreement by tertile of time taken (shortest, medium, longest). Time taken is difficult to interpret: it may decrease with experience, increase when additional care is being taken, or be particularly high when a paper/topic is very complicated or when a study is poorly reported. Spearman's rank correlation (Sedgwick, 2014) was used to test for a trend in inter-rater agreement with time taken. Statistical significance was at the $p < 0.05$ level. If there was discordance, the study was included in the fourth group: missing/discordant time taken.

2.2.3. Themes in user experience

We analysed the qualitative survey data using thematic analysis (Braun & Clarke, 2014). We identified key topics in the user experiences and then the themes within each topic.

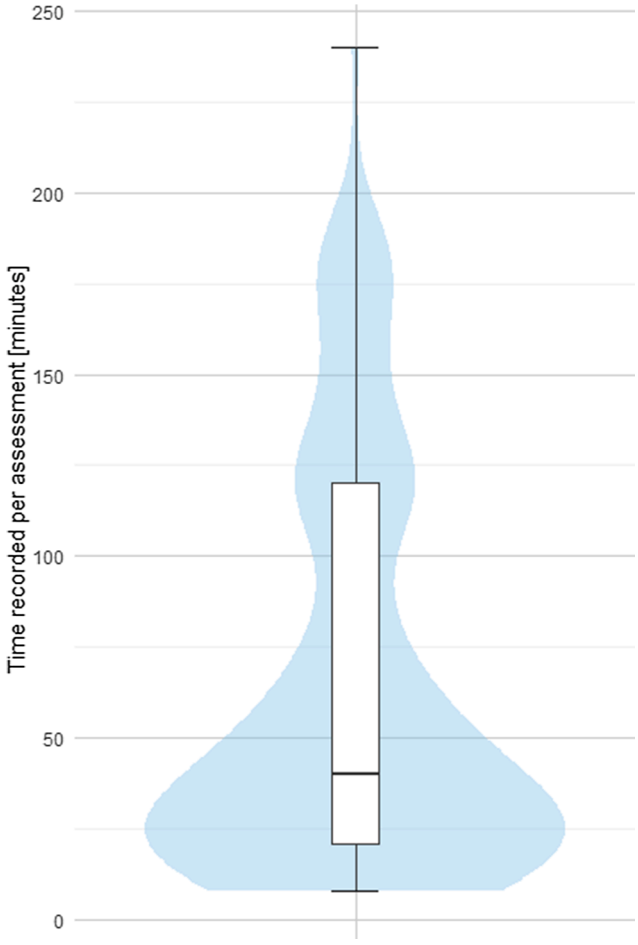


Fig. 1. Time taken as recorded for each study assessment with the RoB-SPEO tool.

3. Results

3.1. Assessor burden

Time recorded for assessments is shown in Fig. 1. Median time taken to assess each study was 40 min (interquartile range 21–120 min). However, the largest proportion of assessors seemed to use approximately 25 min per assessment, while a small proportion of assessors reported spending a very larger number of minutes, such as > 150 min (Fig. 1).

Time taken varied by systematic review, with median times for each systematic review ranging from 20 min and 120 min. Among assessors who carried out ≤ 10 assessments (n = 14), median time reported was 26 min; among those who carried out > 10 assessments (n = 13), median time was 49 min.

3.2. Interrater agreement

3.2.1. Main analyses

Overall inter-rater agreement, as measured with P_i, for each domain of RoB-SPEO is shown in Fig. 2. Agreement ranged from 0.54 (for bias due to differences in numerator and denominator) to 0.82 (for bias due to a lack of blinding of study personnel and selective reporting of exposures).

3.2.2. Subgroup analyses

Inter-rater agreement analysed by subgroups is also shown in Fig. 2. Systematic reviews are anonymised and appear in random order, but colour coding indicates the tertile of raw agreement (white 0.00–0.33, light blue 0.34–0.66, dark blue 0.67–1.00). For Systematic Review A, three domains had raw agreement scores in the middle tertile and five had raw agreement scores in the highest tertile. Systematic Review B in all domains had raw agreement scores in the highest tertile. For Systematic Review C, inter-rater agreement was in the highest tertile for two domains, the middle tertile for five domains, and the lowest tertile for the domain of bias due to differences in numerator and denominator. Agreement scores for Systematic Review D were in the highest tertile across all domains. In summary, inter-rater agreement varied between systematic reviews, which could reflect that agreement is harder to achieve for certain topics.

For 117 studies, all assessors had reviewed the same number of studies included here (≤10 assessments or > 10 assessments). Inter-rater agreement scores by the number of included assessments that assessors had conducted are shown in Fig. 2. Raw agreement was higher for three domains among those pairs whose assessors had each carried out > 10 assessments. This does not appear to support that increased experience would lead to more agreement.

For 128 of the 137 studies, all assessors stated the length of time the assessment took them (271 of the 283 assessments). Time taken was divided into tertiles: ≤25 min, 26–66 min or ≥67 min. For 75 studies,

Table 3
Spearman's rank correlation between time taken (tertiles).

	RoB-SPEO domain	Correlation coefficient	P-value
1	Bias in the selection of participants into the study	0.53	0.000000712***
2	Bias due to lack of blinding of study personnel	0.32	0.0027**
3	Bias due to exposure misclassification	0.34	0.0013**
4	Bias due to incomplete exposure data	0.50	0.00000209***
5	Bias due to selective reporting of exposures	0.26	0.0114*
6	Bias due to conflict of interest	0.47	0.00000915***
7	Bias due to differences in numerator and denominator	0.74	0.0000000000000319***
8	Other bias	0.28	0.0081**

* Statistically significant at the p < 0.05 level
** Statistically significant at the p < 0.01 level
*** Statistically significant at the p < 0.001 level

		Systematic review[a]			Assessor experience with RoB-SPEO[b,c]			Time taken (tertiles)[d,e]			OVERALL[f]
		Systematic review A	Systematic review B	Systematic review C	Systematic review D	All assessors ≤10 assessments (11 records from four systematic reviews)	All assessors >10 assessments (106 records from three systematic reviews)	All assessors recorded ≤25 minutes (18 records from three systematic reviews)	All assessors recorded 26–66 minutes (20 records from three systematic reviews)	All assessors recorded ≥67 (37 records from two systematic review)	
RoB-SPEO domain	(1) Bias in selection of participants into the study	0.34-0.66	≥0.67	0.34-0.66	≥0.67	0.64	0.63	0.59	0.25	1	**0.67**
	(2) Bias due to a lack of blinding of study personnel	≥0.67	≥0.67	≥0.67	≥0.67	0.73	0.86	0.72	0.8	0.97	**0.82**
	(3) Bias due to exposure misclassification	≥0.67	≥0.67	≥0.67	≥0.67	0.54	0.83	0.78	0.67	1	**0.8**
	(4) Bias due to incomplete exposure data	0.34-0.66	≥0.67	0.34-0.66	≥0.67	0.82	0.74	0.56	0.57	0.85	**0.73**
	(5) Bias due to selective reporting of exposures	≥0.67	≥0.67	0.34-0.66	≥0.67	0.73	0.84	0.71	0.9	0.94	**0.82**
	(6) Bias due to conflict of interest	≥0.67	≥0.67	0.34-0.66	≥0.67	0.91	0.8	0.5	0.87	0.97	**0.76**
	(7) Bias due to differences in numerator and denominator	0.34-0.66	≥0.67	<0.33	≥0.67	0.73	0.56	0.18	0.5	1	**0.54**
	(8) Other bias	≥0.67	≥0.67	0.34-0.66	≥0.67	0.91	0.73	0.61	0.76	0.89	**0.74**

Fig. 2. Inter-rater agreement of the RoB-SPEO tool by domain. Footnotes: [a]Data by systematic review has been anonymised and randomized in order. Instead of the colour scale used in the rest of the graph, the scale was split into tertiles and colour coded accordingly, to ensure anonymity (white 0.00-0.33, light blue 0.34-0.66, dark blue 0.67-1.00). [b]Agreement shown for studies where all assessors had carried out a similar number of assessments (≤10 or >10). [c]Domain 1 score missing for one study record, resulting in 105 records from three systematic reviews included in the >10 assessments category; Domain 5 score missing for one study record, resulting in 105 records from three systematic reviews included in the >10 assessments category; Domain 7 score missing for five study records, resulting in 103 records from three systematic reviews included in the >10 assessments category (for the other two with missing scores, the reviewers did not have concordant experience). [d]Agreement shown for studies where all assessors recorded similar time for assessment, and where discordant times were recorded. [e]Domain1 score missing for one study record, resulting in 17 records from three systematic reviews included in the ≤25 minutes category; Domain 5 score missing for one study record, resulting in 17 records from three systematic reviews included in the ≤25 minutes category; Domain 7 score missing for five study records, resulting in 17 records from three systematic reviews included in the ≤25 minutes category, 19 records from three systematic reviews included in the 26-66 minutes category, and 59 records from four systematic reviews included in the discordant times category. [f]Domain 1 and 5 scores missing for one study record, resulting in 137 records from four systematic reviews; Domain 7 score missing for five study records, resulting in 133 records from four systematic reviews.

Table 4
Comparison of inter-rater agreement ratings for Version 4.0 versus Version 6.0 of RoB-SPEO by domain.

	RoB-SPEO domain	Inter-rater reliability, Version 4.0	Inter-rater reliability, Version 6.0	Change from Version 4.0 to Version 6.0
1	Bias in selection of participants into the study	0.33	0.67	+0.34
2	Bias due to a lack of blinding of study personnel	0.65	0.82	+0.17
3	Bias due to exposure misclassification	0.76	0.80	+0.04
4	Bias due to incomplete exposure data	0.31	0.73	+0.42
5	Bias due to selective reporting of exposures	0.80	0.82	+0.02
6	Bias due to conflict of interest	0.51	0.76	+0.25
7	Bias due to differences in numerator and denominator	NA[a]	0.54	NA[a]
8	Other bias	0.51	0.74	+0.23

[a] NA = not applicable (the domain of bias due to differences in numerator and denominator was not included in Version 4.0 of RoB-SPEO).

there was concordance between all assessors regarding reported time taken for the study; for 62 studies, times were missing or discordant between assessors. Inter-rater agreement scores by time taken to review the studies (split by tertile and studies missing time data) are shown in Fig. 2. For all domains, inter-rater agreement was highest for those studies that had the most time dedicated to review them. Raw agreement scores were lowest for studies for which risk of bias assessments had been completed most quickly for seven out of eight domains. For all domains, inter-rater agreement (as measured with Spearman's rank correlation; Sedgwick, 2014) increased significantly with more time taken (Table 3). For four domains, there was a weak correlation between time taken and inter-rater agreement. For three, there was a moderate correlation. For one, bias due to differences in numerator and denominator, there was a strong correlation. Importantly, a test of ordinal ranking by category showed that, for each RoB-SPEO domain, more time taken was correlated with higher inter-rater agreement (with correlation coefficients ranging from 0.26 to 0.74, and all statistically significant [p-values all < 0.05]).

3.3. Comparison to pilot testing results on interrater agreement

Inter-rater agreement assessed after completion of pilot testing of RoB-SPEO Version 4.0 that agreement ranged from 0.31 to 0.80. The differences between inter-rater agreement ratings for Version 4.0 and Version 6.0 are shown in Table 4. For all seven domains included in Version 4.0 of RoB-SPEO, agreement improved for Version 6.0. Agreement for the two domains for which agreement was lowest in Version 4.0 improved the most.

3.4. User experiences with RoB-SPEO

3.4.1. RoB-SPEO's advantages

Assessors described their experiences with using the RoB-SPEO tool. In assessors' descriptions of RoB-SPEO's advantages, we identified three main themes (Box 1, including example quotes from assessors). Assessors described RoB-SPEO as well structured, expressing appreciation for its individual components. They stated that the tool was comprehensive and related well to occupational settings, and they expressed that one key advantage of RoB-SPEO is that it helps establish agreed standards for RoB assessments across systematic reviews of prevalence studies of occupational exposures.

3.4.2. RoB-SPEO's disadvantages

Box 2 displays some example quotes from assessors regarding the three main themes identified as disadvantages of the RoB-SPEO tool by assessors. Some would have liked more detailed descriptions to provide clarity for some of the domains. Additionally, they addressed the external validity of the tool, highlighting challenges applying RoB-SPEO to different studies and suggesting that they believed some of the domains to be less relevant than others. Finally, some comments related to the tool's assessor burden and technical complexity.

Box 1
Themes of assessors' descriptions of RoB-SPEO's advantages.

Theme 1: Structure and components of the RoB-SPEO tool.

(1) RoB-SPEO provides clear descriptions of the bias and criteria for rating.
  (i) *"It is easier to understand the criteria for rating, and write the justification." (Assessor 9)*
  (ii) *"Clear definition, considerations, and criteria for rating." (Assessor 8)*
(2) RoB-SPEO provides useful examples.
  (i) *"Being a tricky domain, it is good that so many examples were provided."* – bias due to exposure misclassification *(Assessor 16)*

Theme 2: External validity of the RoB-SPEO tool.

(1) RoB-SPEO is comprehensive and covers relevant domains.
  (i) *"It relates well to occupational settings." (Assessor 7)*
  (ii) *"Gives a clear and comprehensive tool to assess the risk of bias" (Assessor 3)*
  (iii) *"… it is important to know about numerator and denominator, in order to calculate accurate incidence rate."* – bias due to differences in numerator and denominator *(Assessor 13)*

Theme 3: Standardizing risk of bias assessments across reviews.

(1) RoB-SPEO standardizes risk of bias assessment.
  (i) *"It harmonizes the approach of different evaluators… working in group." (Assessor 1)*

Box 2
Themes of assessors' descriptions of RoB-SPEO's disadvantages.

Theme 1: Structure and components of the RoB-SPEO tool.

(1) Some tool components lack clarity.
  (i) *"Rules are not clearly applicable when the study population is representative for the target population, but exposure groups are not established and the exposure levels measured are not representative for the exposure of the target population…"* – bias due to selection of participants into the study *(Assessor 5)*
  (ii) *"Probably too short description."* – other bias *(Assessor 16)*

Theme 2: External validity.

(1) Difficulties applying RoB-SPEO to studies.
  (i) *"Sometimes it is difficult to evaluate if there are incomplete exposure data. It depends on exposure assessment method."* – related to bias due to incomplete exposure data *(Assessor 6)*
  (ii) *"Probably too standardized; sometimes it would be necessary to take into account differences in study designs, to make it more powerful and meaningful." (Assessor 16)*
(2) Lack of relevance.
  (i) *"Blindness is not always a problem in environmental studies."* –bias due to lack of blinding of study personnel *(Assessor 10)*
  (ii) *"… in quantitative studies… I don't see the risk (apart for analytical mistakes)."* – bias due to exposure misclassification *(Assessor 1)*

Theme 3: Assessor burden.

(1) Workload.
  (i) *"It is usually hard to find numerator or denominator in the article. It is a tough work."* –regarding bias due to differences in numerator and denominator *(Assessor 13)*
  (ii) *"It could be very long to be applied in specific studies…" (Assessor 17)*
(2) Complexity.
  (i) *"The instructions are too complicated to read."* – bias due to selection *(Assessor 17)*
  (ii) *"As an epidemiologist the language is fine but I am not sure if a non-epidemiologist/hygienist might be confused by some of the terms and definitions."* – bias due to misclassification of exposure *(Assessor 8)*

3.4.3. Users' suggestions for improvement

The assessors provided several suggestions regarding improvements that could be made to RoB-SPEO, which we grouped into two themes (Box 3). Some suggested more examples and guidance would be useful and amendments to domains. Additionally, within the theme of assessor burden, there were suggestions to develop an online platform for the tool.

Box 3
Themes of assessors' descriptions of suggestions for improvement of RoB-SPEO.

Theme 1: Structure and components of the RoB-SPEO tool.

(1) Additional examples/guidance for using RoB-SPEO.
  (i) *"Examples. Examples. Examples" (Assessor 3)*
  (ii) *"Giving more indications on how to proceed in case target population could not be identified."* – bias due to selection of participants into the study *(Assessor 17)*
  (iii) *"… there can be very subtle/well hidden source of bias, it would be important to give maybe more examples to help identify it, mostly for the less experienced reviewers."* – other bias *(Assessor 16)*
(2) Suggestions for amending domains of RoB-SPEO.
  (i) *"Only add this domain when the characteristics/designs of the studies… allow to evaluate it in a meaningful way."* – bias due to differences in numerator and denominator *(Assessor 16)*
  (ii) *"The last domain may be not necessary according to my experience…."* – other bias *(Assessor 17)*

Theme 2: Assessor burden.

(1) Platform.
  (i) *"A tool based on Internet or a software could improve the efficiency of assessment." (Assessor 13)*
  (ii) *"It would be good to prepare a tool resident on a website (Dropbox?) and directly linked with other phases of the assessment (data extraction in particular)." (Assessor 1)*

4. Discussion

4.1. Summary of findings

The median time taken to assess a study record using RoB-SPEO was 40 min (with an interquartile range of 21–120 min). Time taken differed substantially between groups of assessors defined by the specific systematic reviews conducted; this may reflect differences in complexity of measuring and assessing the prevalence of occupational exposure to solar ultraviolet radiation, noise, dusts/fibres, and ergonomic factors.

Using a raw measure of agreement (*Pi*), we found that overall inter-rater agreement varied by domain from 0.54 (for bias due to differences in numerator and denominator) to 0.82 (for bias due to lack of blinding of study personnel and selective reporting of exposures). This suggests good agreement across domains. The lowest agreement, which was for the domain of bias due to differences in numerator and denominator, seemed to be driven by discordance where one of the assessors for studies rated this domain as "unclear" or "not possible to determine"; this may suggest that this bias (specific to prevalence measures) may not be as well understood as other biases covered in RoB-SPEO.

Agreement varied by systematic review, again perhaps due to exposure-specific differences in complexity of assessing RoB in studies. Variation was also seen with differing assessor experience with the RoB-SPEO tool; however, we would have expected more experienced assessors to have had higher inter-rater agreement, which was not the case across domains. Furthermore, an important finding of our performance testing study is that agreement increased with more time taken to carry out an assessment.

Users described RoB-SPEO advantages comprising the clarity of its overall structure and components, its comprehensiveness and coverage of relevant RoB domains, and that it enabled assessors to standardize their assessments, both within and across systematic reviews. The users described disadvantages including that some tool components lacked clarity, they encountered difficulties applying RoB-SPEO to some studies, and a perceived lack of relevance, plus concerns related to assessor burden in terms of managing workload and complexity. Assessors suggested improvement of RoB-SPEO could focus on further attention to structure and components of the RoB-SPEO tool, primarily adding further examples/guidance for using RoB-SPEO and amending some RoB-SPEO domains. Some users proposed that an online platform for conducting RoB-SPEO assessments could be created to reduce assessor burden. This suggestion could give more flexibility and be useful in addressing the sometimes conflicting opinions of users regarding the necessary level of guidance and number of examples. An online tool could allow provision of more instructions/examples for assessors who need it, for example it would be possible for users to expand or click through to additional text if they were less experienced or required further guidance. It could also provide guidance on how to rate domains that they feel are less relevant (e.g. if a domain is less relevant to a certain topic or study design).

While assessors are required to reach consensus on a final rating for each domain for each study, after making their individual ratings using RoB-SPEO, such a tool should provide a reliable measure of the risk of bias. Inter-rater agreement is way to assess reliability; therefore, despite the role of individual judgement, variability in individual risk of bias ratings between users would ideally be low, even if it cannot be expected to be zero.

The higher level of agreement observed for Version 6.0 compared to Version 4.0 of the RoB-SPEO tool may suggest that the further development of the tool and its guidance notes improved inter-rater agreement. Alternatively, or in addition, experience with the tool may have improved between-rater agreement in their ratings. Some assessors included in this study were involved in piloting Version 4.0 of RoB-SPEO, so they had prior experience with (an earlier version of) the tool. Most if not all assessors received training and, subsequently, customized guidance from WHO and ILO on the RoB-SPEO tool. Jeyaraman et al. (2020b) found inter-rater reliability of the tools Risk of Bias in Non-Randomized Studies of Exposures (ROB-NRSE) and Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I) improved after training and customized guidance.

The domain of bias due to difference in numerator and denominator was introduced to the RoB-SPEO tool relatively late, so guidance for this domain could perhaps benefit from further testing and refinement, especially since our current study found that assessors most commonly provided missing ratings for bias in this domain.

We have made the anonymised raw data (ratings) available open access in this article (Appendix 2); future studies can re-analyse these same data using different methods and ratings for inter-rater agreement (if and when of interest).

4.2. Comparison to assessor burden and measures of inter-rater agreement of other tools

4.2.1. Comparison to assessor burden of other tools

Assessor burden of the RoB-SPEO (median 40 min) appears to be comparative with that of other RoB tools. A recent performance testing study of the ROBINS-I and ROBINS-E tools found that the average time taken to conduct an assessment with the tools was 42.7 $\pm$ 7.7 min and 48 $\pm$ 8.3 min, respectively (Jeyaraman et al., 2020a). Eick et al. (2020), in their study comparing three different RoB assessment tools, found that the average time needed for the assessment was 20 min per study with the OHAT tool, 32 min with the IRIS tool, and 40 min with the TSCA tool. However, there was a large range in the times reported for assessments with RoB-SPEO (interquartile range 21–120).

4.2.2. Comparison to inter-rater agreement of other tools

That studies assessing inter-rater agreement and reliability of RoB tools use different ratings makes it difficult to directly compare results across such studies. Studies of the ROBINS-I and Cochrane ROB 2.0 tools, for example, categorized kappa values (different to our measure of P_i) (Minozzi et al., 2020; Couto et al., 2015).

4.3. Limitations of this study and potential further development of RoB-SPEO

We did not receive RoB-SPEO ratings for all of the data we requested for all studies. One systematic review had not completed study selection, some assessments were not returned for some included studies, and some individual assessments missed ratings for one or more domains. There may be differences between the responses we received and those we did not.

Variation in reports of time taken could have been due to assessors quantifying this differently. The reporting form asked "Time needed for the assessment"; some may have included time taken to read the paper, whereas others may not have. Future requests to report time taken should clarify how this should be quantified. Additionally, the performance of RoB-SPEO could be tested on systematic reviews on additional topics, for example those relating to exposures to biological or psychosocial occupational risk factors to human health. This is particularly important if complexity of topic can affect inter-rater agreement and assessor burden. RoB-SPEO's performance could be tested when the tool is used in combination with other tools or approaches for occupational exposure prevalence studies, such as the QoE-SPEO approach for assessing quality of evidence in bodies of evidence from such studies, which comprises assessments at the level of the entire body of evidence of risk of bias, as well as of inconsistency, indirectness, imprecision and publication bias (Pega et al., in preparation).

Further to this, while our study includes several assessors, as recommended by Pieper et al (Pieper et al. 2017), we were not able to account for differences between the 27 assessors. We attempted to consider assessor experience with the tool when calculating inter-rater agreement, but many diverse factors could affect ratings, including

experience with this type of assessment, expertise regarding the topic under review and engagement with the project. Previous studies have highlighted differences between reviewers, for example in how they interpret questions (Gates et al. 2020); that coding behaviours vary both between and within individuals over time (Belur et al. 2018); and that pairs of reviewers who have previously worked together demonstrated inter-rater reliability (Pieper et al., 2019), which we did not have information on.

This study was undertaken opportunistically, using assessments carried out as part of the WHO/ILO Joint Estimates systematic reviews of prevalence. The assessments were not carried out as part of a study that was designed specifically to test the tool and there was no pre-specified protocol for this investigation of the RoB-SPEO tool. However, this study expands on the analysis carried out previously (Pega et al., 2020b) and uses the exact same methods. Additionally, it efficiently made use of data that were already generated as part of systematic reviews conducted for the WHO/ILO Joint Estimates (Pega et al., 2021a,b, World Health Organization and International Labour Organization, 2021a,b) by a large number of diverse assessors from across the globe (Li et al., 2018; Descatha et al., 2018; Godderis et al., 2018; Bonafede et al., 2021; Descatha et al., 2020; Hulshof et al., 2019; Li et al., 2020; Mandrioli et al., 2018; Pachito et al., 2021; Paulo et al., 2019; Pega et al., 2020b; Teixeira et al., 2019, 2021b; Tenkate et al., 2019; Hulshof et al., 2021a; Hulshof et al., 2021b; Rugulies et al., 2019; Teixeira et al., 2021a; Schlünssen et al., in preparation). Additional testing in a different format would also be a useful exercise to assess inter-rater agreement. For example, a large number of assessors could be asked to assess a smaller number of studies using RoB-SPEO. Further, our assessment does not test for other forms of reliability or validity, or other contexts of evaluation, which could be considered in the future. Future developments, based on the feedback received from RoB-SPEO users, may help to improve the tool further in terms of reducing assessor burden and improving inter-rater agreement.

5. Conclusions

Our study suggests that risk of bias assessments conducted with the RoB-SPEO tool place a similar time burden on assessors as do some other RoB tools. Overall, for all eight domains of the RoB-SPEO tool, raters achieved raw agreement scores between 0.54 and 0.82, indicating good agreement across domains. Additional time spent on assessments may improve inter-rater agreement. It is likely that agreement varies with systematic review topic, which may reflect variations in difficulty measuring different exposures or in risk of bias within different systematic review topics. Further training and development of guidance notes is likely to be useful for the domains of bias which had the relatively lowest inter-rater reliability (i.e., particularly for bias due to difference in numerator and denominator) to further develop the RoB-SPEO tool.

Sponsors

The sponsors of this analysis are the World Health Organization and the International Labour Organization.

Author contribution

Conceptualization: FP. Data curation: NCM, DTCdS, TL, DM, AM, MSP, PS, VS, LRT, FP. Formal analysis: NCM, KNS, FP. Funding acquisition: FP. Investigation: NCM, FP. Methodology: NCM, DG, DM, FP. Project administration: FP. Software: NCM, KNS. Supervision: FP. Validation: All authors. Visualization: NCM, KNS, FP. Writing – original draft: NCM, FP. Writing – review & editing: all authors.

Funding sources

This study was prepared with financial support to the World Health Organization from the National Institute for Occupational Safety and Health of the Centres for Disease Control and Prevention of the United States of America (Grant 1E11OH0010676-02; Grant 6NE11OH010461-02-01; and Grant 5NE11OH010461-03-00); the German Federal Ministry of Health (BMG Germany) under the BMG-WHO Collaboration Programme 2020–2023 (WHO specified award ref. 70672); and the Spanish Agency for International Cooperation (AECID) (WHO specified award ref. 71208). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr Paul Whaley (Associate Editor for Systematic Reviews, *Environment International*; Lancaster Environment Centre, Lancaster University, United Kingdom) for the editorial guidance and support. We thank Professor Carel T.J. Hulshof for providing data from risk of bias assessments for one systematic review. We thank Dr Yuka Ujita (ILO) for comments on an earlier version of the manuscript.

Disclaimer

The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.107005.

References

Armijo-Olivo, S., Stiles, C.R., Hagen, N.A., Biondo, P.D., Cummings, G.G., 2012. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J. Eval. Clin. Pract. 18 (1), 12–18.

World Health Organization, International Labour Organization, 2021b. WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury, 2000-2016: Technical Report with Data Sources and Methods. World Health Organization, International Labour Organization, Geneva, Switzerland.

World Health Organization, International Labour Organization, 2021a. WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury, 2000-2016: Global Monitoring Report. World Health Organization, International Labour Organization, Geneva, Switzerland.

Belur, J., Tompson, L., Thornton, A., Simon, M., 2018. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociol. Methods Res. 50(2), pp 837–865. doi: 10.1177/0049124118799372.

Bilandzic, A., Fitzpatrick, T., Rosella, L., Henry, D., 2016. Risk of bias in systematic reviews of non-randomized studies of adverse cardiovascular effects of thiazolidinediones and cyclooxygenase-2 inhibitors: application of a new cochrane risk of bias tool. PLoS Med. 13(4), pp. e1001987.

Boutron, I., Page, M.J., Higgins, J.P.T., Altman, D.G., Lundh, A., Hróbjartsson, A., 2020. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V. A. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions version 6.1: Cochrane.

Rugulies, R., Sørensen, K., Di Tecco, C., Bonafede, M., Rondinone, B.M., Ahn, S., Ando, E., Ayuso-Mateos, J.L., Cabello, M., Descatha, A., Dragano, N., Durand-Moreau, Q., Eguchi, H., Gao, J., Godderis, L., Kim, J., Li, J., Madsen, I.E.H., Pachito, D.V., Sembajwe, G., Siegrist, J., Tsuno, K., Ujita, Y., Wang, J.L., Zadow, A., Iavicoli, S., Pega, F., 2021. The effect of exposure to long working hours on depression: A systematic review and meta-analysis from the WHO/ILO Joint

Estimates of the Work-Related Burden of Disease and Injury. Environ. Int. 155, 106629. https://doi.org/10.1016/j.envint.2021.106629.
Braun, V., Clarke, V., 2014. What can "thematic analysis" offer health and wellbeing researchers? Int. J. Qual. Stud. Health Well-being 16 (9), 26152. https://doi.org/10.3402/qhw.v9.26152.
Couto, E., Pike, E., Torkilseng, E.B., Klemp, M., 2015. Inter-rater reliability of the Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI). In: Filtering the information overload for better decisions. Abstracts of the 23rd Cochrane Colloquium. John Wiley & Sons, Vienna, Austria.
Descatha, A., Sembajwe, G., Baer, M., Boccuni, F., Di Tecco, C., Duret, C., Evanoff, B.A., Gagliardi, D., Ivanov, I.D., Leppink, N., Marinaccio, A., Magnusson Hanson, L.L., Ozguler, A., Pega, F., Pell, J., Pico, F., Prüss-Üstün, A., Ronchetti, M., Roquelaure, Y., Sabbath, E., Stevens, G.A., Tsutsumi, A., Ujita, Y., Iavicoli, S., 2018. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on stroke. Environ. Int. 119, 366–378.
Descatha, A., Sembajwe, G., Pega, F., Ujita, Y., Baer, M., Boccuni, F., Di Tecco, C., Duret, C., Evanoff, B.A., Gagliardi, D., Godderis, L., Kang, S.-K., Kim, B.J., Li, J., Magnusson Hanson, L.L., Marinaccio, A., Ozguler, A., Pachito, D., Pell, J., Pico, F., Ronchetti, M., Roquelaure, Y., Rugulies, R., Schouteden, M., Siegrist, J., Tsutsumi, A., Iavicoli, S., 2020. The effect of exposure to long working hours on stroke: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 142, 105746. https://doi.org/10.1016/j.envint.2020.105746.
Eick, S.M., Goin, D.E., Chartres, N., Lam, J., Woodruff, T.J., 2020. Assessing risk of bias in human environmental epidemiology studies using three tools: different conclusions from different tools. Syst. Rev. 9 (1), 249.
Gates, M., Gates, A., Duarte, G., Cary, M., Becker, M., Prediger, B., Vandermeer, B., Fernandes, R.M., Pieper, D., Hartling, L., 2020. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. J. Clin. Epidemiol. 125, 9–15.
Godderis, L., Boonen, E., Cabrera Martimbianco, A.L., Delvaux, E., Ivanov, I.D., Lambrechts, M.C., Latorraca, C.O.C., Leppink, N., Pega, F., Prüss-Ustün, A.M., Riera, R., Ujita, Y., Pachito, D.V., 2018. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on alcohol consumption and alcohol use disorders. Environment International 120, 22–33. https://doi.org/10.1016/j.envint.2018.07.025.
Gwet, K., 2001. (2001) Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters. STATAXIS Publishing Company, Gaithersburg, MD.
Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61 (Pt 1), 29–48.
Hartling, L., Hamm, M.P., Milne, A., Vandermeer, B., Santaguida, P.L., Ansari, M., Tsertsvadze, A., Hempel, S., Shekelle, P., Dryden, D.M., 2013. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J. Clin. Epidemiol. 66 (9), 973–981.
Hoy, D., Brooks, P., Woolf, A., Blyth, F., March, L., Bain, C., Baker, P., Smith, E., Buchbinder, R., 2012. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. J. Clin. Epidemiol. 65 (9), 934–939.
Hulshof, C.T.J., Colosio, C., Daams, J.G., Ivanov, I.D., Prakash, K.C., Kuijer, P.P.F.M., Leppink, N., Mandic-Rajcevic, S., Masci, F., van der Molen, H.F., Neupane, S., Nygård, C.-H., Oakman, J., Pega, F., Proper, K., Prüss-Üstün, A.M., Ujita, Y., Frings-Dresen, M.H.W., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to occupational ergonomic risk factors and of the effect of exposure to occupational ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases. Environ. Int. 125, 554–566.
Hulshof, C.T.J., Pega, F., Neupane, S., van der Molen, H.F., Colosio, C., Daams, J.G., Descatha, A., Kc, P., Kuijer, P., Mandic-Rajcevic, S., Masci, F., Morgan, R.L., Nygard, C.H., Oakman, J., Proper, K.I., Solovieva, S., Frings-Dresen, M.H.W., 2021a. The prevalence of occupational exposure to ergonomic risk factors: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 146, pp. 106157.
Hulshof, C.T.J., Pega, F., Neupane, S., Colosio, C., Daams, J.G., Kc, P., Kuijer, P., Mandic-Rajcevic, S., Masci, F., van der Molen, H.F., Nygard, C.H., Oakman, J., Proper, K.I., Frings-Dresen, M.H.W., 2021b. The effect of occupational exposure to ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int., pp. 106349.
Jeyaraman, M.M., Al-Yousif, N., Robson, R.C., Copstein, L., Balijepalli, C., Hofer, K., Fazeli, M.S., Ansari, M.T., Tricco, A.C., Rabbani, R., Abou-Setta, A.M., 2020a. Inter-rater reliability and validity of risk of bias instrument for non-randomized studies of exposures: a study protocol. Syst. Rev. 9 (1), 32.
Jeyaraman, M.M., Robson, R.C., Pollock, M., Copstein, L., Balijepalli, C., Hofer, K., Xia, J., Al-Yousif, N., Mansour, S., Fazeli, M.S., Ansari, M.T., Tricco, A.C., Rabbani, R., Abou-Setta, A.M., 2020b. Impact of training and guidance on the inter-rater and inter-consensus reliability of risk of bias instruments for non-randomized studies. Advances in Evidence Synthesis: special issue Cochrane Database of Systematic Reviews.
Kennedy, C.E., Fonner, V.A., Armstrong, K.A., Denison, J.A., Yeh, P.T., O'Reilly, K.R., Sweat, M.D., 2019. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Syst. Rev. 8 (1), 3.
Krauth, D., Woodruff, T.J., Bero, L., 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ. Health Perspect. 121 (9), 985–992.
Li, J., Brisson, C., Clays, E., Ferrario, M.M., Ivanov, I.D., Landsbergis, P., Leppink, N., Pega, F., Pikhart, H., Prüss-Üstün, A., Rugulies, R., Schnall, P.L., Stevens, G., Tsutsumi, A., Ujita, Y., Siegrist, J., 2018. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on ischaemic heart disease'. Environ. Int. 119, 558–569.
Li, J., Pega, F., Ujita, Y., Brisson, C., Clays, E., Descatha, A., Ferrario, M.M., Godderis, L., Iavicoli, S., Landsbergis, P.A., Metzendorf, M.-I., Morgan, R.L., Pachito, D.V., Pikhart, H., Richter, B., Roncaioli, M., Rugulies, R., Schnall, P.L., Sembajwe, G., Trudel, X., Tsutsumi, A., Woodruff, T.J., Siegrist, J., 2020. The effect of exposure to long working hours on ischaemic heart disease: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 142, 105739. https://doi.org/10.1016/j.envint.2020.105739.
Losilla, J.-M., Oliveras, I., Marin-Garcia, J.A., Vives, J., 2018. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J. Clin. Epidemiol. 101, 61–72.
Mandrioli, D., Schlünssen, V., Ádám, B., Cohen, R.A., Colosio, C., Chen, W., Fischer, A., Godderis, L., Göen, T., Ivanov, I.D., Leppink, N., Mandic-Rajcevic, S., Masci, F., Nemery, B., Pega, F., Prüss-Üstün, A., Sgargi, D., Ujita, Y., van der Mierden, S., Zungu, M., Scheepers, P.T.J., 2018. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of occupational exposure to dusts and/or fibres and of the effect of occupational exposure to dusts and/or fibres on pneumoconiosis. Environ. Int. 119, 174–185.
Mandrioli, D., Silbergeld, E.K., 2016. Evidence from toxicology: the most essential science for prevention. Environ. Health Perspect. 124 (1), 6–11.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. (Zagreb) 22 (3), 276–282.
Minozzi, S., Cinquini, M., Gianola, S., Gonzalez-Lorenzo, M., Banzi, R., 2020. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J. Clin. Epidemiol. 126, 37–44.
Morgan, R.L., Thayer, K.A., Santesso, N., Holloway, A.C., Blain, R., Eftim, S.E., Goldstone, A.E., Ross, P., Ansari, M., Akl, E.A., Filippini, T., Hansell, A., Meerpohl, J. J., Mustafa, R.A., Verbeek, J., Vinceti, M., Whaley, P., Schünemann, H.J., 2019. A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. Environ. Int. 122, 168–184.
Morgan, R.L., Thayer, K.A., Santesso, N., Holloway, A.C., Blain, R., Eftim, S.E., Goldstone, A.E., Ross, P., Guyatt, G., Schünemann, H.J., 2018a. Evaluation of the risk of bias in non-randomized studies of interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: Rationale and preliminary instrument development. Environ. Int. 120, 382–387.
Morgan, R.L., Whaley, P., Thayer, K.A., Schunemann, H.J., 2018b. Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ. Int. 121 (Pt 1), 1027–1031.
Munn, Z., Moola, S., Lisy, K., Riitano, D., Tufanaru, C., 2015. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int. J. Evid. Based Healthc. 13 (3), 147–153.
Munn, Z., Moola, S., Riitano, D., Lisy, K., 2014. The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. Int. J. Health Policy Manage. 3 (3), 123–128.
NHMRC, 2019. Guidelines for Guidelines: Assessing risk of bias: NHMRC. Available at: https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias.
NTP (National Toxicology Program), 2016. Report on Carcinogens, Fourteenth Edition, Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service. Available at: https://ntp.niehs.nih.gov/go/roc14.
Pachito, D.V., Pega, F., Bakusic, J., Boonen, E., Clays, E., Descatha, A., Delvaux, E., De Bacquer, D., Koskenvuo, K., Kröger, H., Lambrechts, M.C., Latorraca, C.O.C., Li, J., Cabrera Martimbianco, A.L., Riera, R., Rugulies, R., Sembajwe, G., Siegrist, J., Sillanmäki, L., Sumanen, M., Suominen, S., Ujita, Y., Vandersmissen, G., Godderis, L., 2021. The effect of exposure to long working hours on alcohol consumption, risky drinking and alcohol use disorder: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 146, 106205. https://doi.org/10.1016/j.envint.2020.106205.
Paulo, M.S., Adam, B., Akagwu, C., Akparibo, I., Al-Rifai, R.H., Bazrafshan, S., Gobba, F., Green, A.C., Ivanov, I., Kezic, S., Leppink, N., Loney, T., Modenese, A., Pega, F., Peters, C.E., Prüss-Üstün, A.M., Tenkate, T., Ujita, Y., Wittlich, M., John, S.M., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of occupational exposure to solar ultraviolet radiation and of the effect of occupational exposure to solar ultraviolet radiation on melanoma and non-melanoma skin cancer. Environ. Int. 126, 804–815.
Pega, Frank, Chartres, Nicholas, Guha, Neela, Modenese, Alberto, Morgan, Rebecca L., Martínez-Silveira, Martha S., Loomis, Dana, 2020a. The effect of occupational exposure to welding fumes on trachea, bronchus and lung cancer: A protocol for a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 145, 106089. https://doi.org/10.1016/j.envint.2020.106089.
Pega, Frank, Hamzaoui, Halim, Náfrádi, Bálint, Momen, Natalie C., 2021b. Global, regional and national burden of disease attributable to 19 selected occupational risk factors for 183 countries, 2000-2016: A systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Scandinavian Journal of Work, Environment & Health. https://doi.org/10.5271/sjweh.4001. In press.

Pega, Frank, Momen, Natalie C., Ujita, Yuka, Driscoll, Tim, Whaley, Pauly, 2021c. Systematic reviews and meta-analyses for the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 155, 106605 https://doi.org/10.1016/j.envint.2021.106605.

Pega, Frank, Náfrádi, Bálint, Momen, Natalie C., Ujita, Yuka, Streicher, Kain N., Prüss-Üstün, Annette M., Descatha, Alexis, Driscoll, Tim, Fischer, Frida M., Godderis, Lode, Kiiver, Hannah M., Li, Jian, Magnusson Hanson, Linda L., Rugulies, Reiner, Sørensen, Kathrine, Woodruff, Tracey J., 2021a. Global, regional, and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000-2016: A systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environment International 154, 106595. https://doi.org/10.1016/j.envint.2021.106595.

Pega, F., Norris, S.L., Backes, C., Bero, L.A., Descatha, A., Gagliardi, D., Godderis, L., Loney, T., Modenese, A., Morgan, R.L., Pachito, D., Paulo, M.B.S., Scheepers, P.T.J., Schlünssen, V., Sgargi, D., Silbergeld, E.K., Sorensen, K., Sutton, P., Tenkate, T., Correa, Torreao, da Silva, D., Ujita, Y., van Deventer, E., Woodruff, T.J., Mandrioli, D., 2020b. RoB-SPEO: A tool for assessing risk of bias in studies estimating the prevalence of exposure to occupational risk factors from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int. 135, 105039.

Pega, F., Gagliardi, D., Bero, L., Boccuni, F., Chartres, N., Descatha, A., Godderis, L., Loney, T., Mandrioli, D., Modenese, A., Morgan, R., Pachito, D., Paulo, M., Scheepers, P., Tenkate, T., Norris, S., in preparation. QoE-SPEO: An approach for assessing the quality of evidence in studies estimating prevalence of exposure to occupational risk factors from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int.

Pieper, D., Jacobs, A., Weikert, B., Fishta, A., Wegewitz, U., 2017. Inter-rater reliability of AMSTAR is dependent on the pair of reviewers. BMC Med. Res. Method. 17 (1), 98.

Pieper, Dawid, Puljak, Livia, González-Lorenzo, Marien, Minozzi, Silvia, 2019. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J. Clin. Epidemiol. 108, 26–33.

Porta, M., 2014. A dictionary of epidemiology, 6 ed. Oxford University Press, New York, NY.

Rooney, A.A., Cooper, G.S., Jahnke, G.D., Lam, J., Morgan, R.L., Boyles, A.L., Ratcliffe, J. M., Kraft, A.D., Schunemann, H.J., Schwingl, P., Walker, T.D., Thayer, K.A., Lunn, R. M., 2016. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ. Int. 92–93, pp. 617–629.

Rugulies, R., Ando, E., Ayuso-Mateos, J.L., Bonafede, M., Cabello, M., Di Tecco, C., Dragano, N., Durand-Moreau, Q., Eguchi, H., Gao, J., Garde, A.H., Iavicoli, S., Ivanov, I.D., Leppink, N., Madsen, I.E.H., Pega, F., Pruss-Ustun, A.M., Rondinone, B. M., Sorensen, K., Tsuno, K., Ujita, Y., Zadow, A., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to long working hours and of the effect of exposure to long working hours on depression. Environ. Int. 125, 515–528.

Savovic, J., Weeks, L., Sterne, J.A., Turner, L., Altman, D.G., Moher, D., Higgins, J.P., 2014. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. Syst. Rev. 3, pp. 37.

Schlünssen, V., Mandrioli, D., Pega, F., Adam, B., Chen, W., Cohen, R.A., Colosio, C., Godderis, L., Goen, T., Hadkhale, K., Kunpeuk, W., Lou, J., Mandic-Rajcevic, S., Masci, F., Nemery, B., Popa, M., Rajatanavin, N., Siriruttanapruk, S., Sun, X., Suphanchaimat, R., Thammawijaya, P., Sgargi, D., Ujita, Y., van der Mierden, S., Vangelova, K., Ye, M., Zungu, M., Scheepers, P.T.J., in preparation. The prevalences and levels of occupational exposure to dusts and/or fibres (silica, asbestos and coal): A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. Environ. Int.

Sedgwick, P., 2014. Spearman's rank correlation coefficient. BMJ 349, g7327.

Tenkate, T., Adam, B., Al-Rifai, R.H., Chou, B.R., Gobba, F., Ivanov, I.D., Leppink, N., Loney, T., Pega, F., Peters, C.E., Prüss-Üstün, A.M., Silva Paulo, M., Ujita, Y., Wittlich, M., Modenese, A., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of occupational exposure to solar ultraviolet radiation and of the effect of occupational exposure to solar ultraviolet radiation on cataract. Environ. Int. 125, 542–553.

Teixeira, L.R., Azevedo, T.M., Bortkiewicz, A., Corrêa da Silva, D.T., de Abreu, W., de Almeida, M.S., de Araujo, M.A.N., Gadzicka, E., Ivanov, I.D., Leppink, N., Macedo, M.R.V., de S. Maciel, E.M.G., Pawlaczyk-Łuszczyńska, M., Pega, F., Prüss-Üstün, A.M., Siedlecka, J., Stevens, G.A., Ujita, Y., Braga, J.U., 2019. WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to occupational noise and of the effect of exposure to occupational noise on cardiovascular disease. Environ. Int. 125, 567–578.

Teixeira, L.R., Pega, F., de Abreu, W., de Almeida, M.S., de Andrade, C.A., Azevedo, T.M., Dzhambov, A.M., Hu, W., Macedo, M.R.V., Martinez-Silveira, M.S., Sun, X., Zhang, M., Zhang, S., Correa da Silva, D.T., 2021a. The prevalence of occupational exposure to noise: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury.

Teixeira, Liliane R., Pega, Frank, Dzhambov, Angel M., Bortkiewicz, Alicja, da Silva, Denise T. Correa, de Andrade, Carlos A.F., Gadzicka, Elzbieta, Hadkhale, Kishor, Iavicoli, Sergio, Martínez-Silveira, Martha S., Pawlaczyk-Łuszczyńska, Małgorzata, Rondinone, Bruna M., Siedlecka, Jadwiga, Valenti, Antonio, Gagliardi, Diana, 2021b. The effect of occupational exposure to noise on ischaemic heart disease, stroke and hypertension: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-Related Burden of Disease and Injury. Environ. Int. 154, 106387. https://doi.org/10.1016/j.envint.2021.106387.

The Joanna Briggs Institute, 2017. JBI critical appraisal checklist for studies reporting prevalence data. Available at: http://joannabriggs.org/research/criticalappraisal-tools.html.

Vandenberg, L.N., Agerstrand, M., Beronius, A., Beausoleil, C., Bergman, A., Bero, L.A., Bornehag, C.G., Boyer, C.S., Cooper, G.S., Cotgreave, I., Gee, D., Grandjean, P., Guyton, K.Z., Hass, U., Heindel, J.J., Jobling, S., Kidd, K.A., Kortenkamp, A., Macleod, M.R., Martin, O.V., Norinder, U., Scheringer, M., Thayer, K.A., Toppari, J., Whaley, P., Woodruff, T.J., Ruden, C., 2016. 'A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environ. Health 15(1), pp. 74.

Whaley, P., Letcher, R.J., Covaci, A., Alcock, R., 2016. Raising the standard of systematic reviews published in Environment International. Environ. Int. 97, 274–276.

Williams, G.M., Najman, J.M., Clavarino, A., 2006. Correcting for numerator/denominator bias when assessing changing inequalities in occupational class mortality, Australia 1981–2002. Bull. World Health Organ. 84 (3), 198–203.

Woodruff, T.J., Sutton, P., 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ. Health Perspect. 122 (10), 1007–1014.